

Deep Self-Convolutional Activations Descriptor for Dense Cross-Modal Correspondence

Seungryong Kim^{1*}, Dongbo Min², Stephen Lin³, and Kwanghoon Sohn¹

¹Yonsei University, ²Chungnam National University, ³Microsoft Research

Abstract. We present a novel descriptor, called deep self-convolutional activations (DeSCA), designed for establishing dense correspondences between images taken under different imaging modalities, such as different spectral ranges or lighting conditions. Motivated by descriptors based on local self-similarity (LSS), we formulate a novel descriptor by leveraging LSS in a deep architecture, leading to better discriminative power and greater robustness to non-rigid image deformations than state-of-the-art cross-modality descriptors. The DeSCA first computes self-convolutions over a local support window for randomly sampled patches, and then builds self-convolution activations by performing an average pooling through a hierarchical formulation within a deep convolutional architecture. Finally, the feature responses on the self-convolution activations are encoded through a spatial pyramid pooling in a circular configuration. In contrast to existing convolutional neural networks (CNNs) based descriptors, the DeSCA is training-free (i.e., randomly sampled patches are utilized as the convolution kernels), is robust to cross-modal imaging, and can be densely computed in an efficient manner that significantly reduces computational redundancy. The state-of-the-art performance of DeSCA on challenging cases of cross-modal image pairs is demonstrated through extensive experiments.

1 Introduction

In many computer vision and computational photography applications, images captured under different imaging modalities are used to supplement the data provided in color images. Typical examples of other imaging modalities include near-infrared [1,2,3] and dark flash [4] photography. More broadly, photos taken under different imaging conditions, such as different exposure settings [5], blur levels [6,7], and illumination [8], can also be considered as cross-modal [9,10].

Establishing dense correspondences between cross-modal image pairs is essential for combining their disparate information. Although powerful global optimizers may help to improve the accuracy of correspondence estimation to some extent [11,12], they face inherent limitations without help of suitable matching descriptors [13]. The most popular local descriptor is scale invariant feature transform (SIFT) [14], which provides relatively good matching performance

* This work is done while Seungryong Kim was an intern at Microsoft Research.

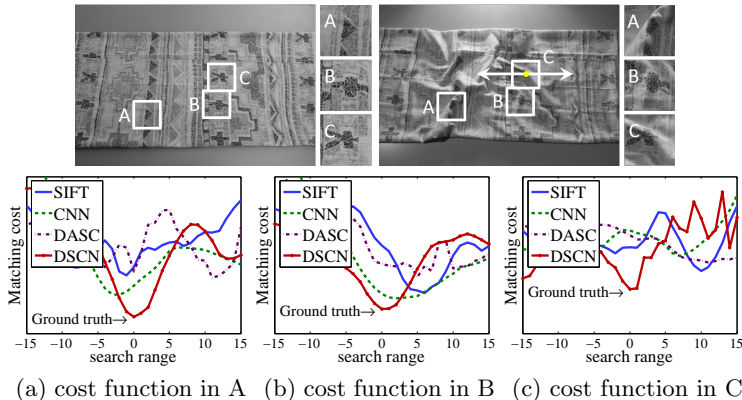


Fig. 1. Examples of matching cost profiles, computed with different descriptors along the scan lines of A, B, and C for image pairs under severe non-rigid deformations and illumination changes. Unlike other descriptors, DeSCA yields reliable global minimum.

when there are small photometric variations. However, conventional descriptors such as SIFT often fail to capture reliable matching evidences in cross-modal image pairs due to their different visual properties [9,10].

Recently, convolutional neural networks (CNNs) based features [15,16,17,18,19] have emerged as a robust alternative with high discriminative power. However, CNN-based descriptors cannot satisfactorily deal with severe cross-modality appearance differences, since they use shared convolutional kernels across images which lead to inconsistent responses similar to conventional descriptors [19,20]. Furthermore, they do not scale well for dense correspondence estimation due to their high computational complexity. Though recent works [21] propose an efficient method that extracts dense outputs through the deep CNNs, they do not extract dense CNN features for all pixels individually. More seriously, their methods were usually designed to perform a specific task only, *e.g.*, semantic segmentation, not to provide a general purpose descriptor like ours.

To address the problem of cross-modal appearance changes, feature descriptors have been proposed based on local self-similarity (LSS) [22], which is motivated by the notion that the geometric layout of local internal self-similarities is relatively insensitive to imaging properties. The state-of-the-art descriptor for cross-modal dense correspondence, called dense adaptive self-correlation (DASC) [10], makes use of LSS and has demonstrated high accuracy and speed on cross-modal image pairs. However, DASC suffers from two significant shortcomings. One is its limited discriminative power due to a limited set of patch sampling patterns used for modeling internal self-similarities. In fact, the matching performance of DASC may fall well short of CNN-based descriptors on images that share the same modality. The other major shortcoming is that the DASC descriptor does not provide the flexibility to deal with non-rigid deformations, which leads to lower robustness in matching.

In this paper, we introduce a novel descriptor, called deep self-convolutional activations (DeSCA), that overcomes the shortcomings of DASC while providing

dense cross-modal correspondences. This work is motivated by the observation that local self-similarity can be formulated in a deep convolutional architecture to enhance discriminative power and gain robustness to non-rigid deformations. Unlike the DASC descriptor that selects patch pairs within a support window and calculates the self-similarity between them, we compute self-convolutional activations that more comprehensively encode the intrinsic structure by calculating the self-similarity between randomly selected patches and all of the patches within the support window. These self-convolutional responses are aggregated through spatial pyramid pooling in a circular configuration, which yields a representation less sensitive to non-rigid image deformations than the fixed patch selection strategy used in DASC. To further enhance the discriminative power and robustness, we build hierarchical self-convolutional layers resembling a deep architecture used in CNN, together with nonlinear and normalization layers. For efficient computation of DeSCA over densely sampled pixels, we calculate the self-convolutional activations through fast edge-aware filtering.

DeSCA resembles a CNN in its deep, multi-layer, and convolutional structure. In contrast to existing CNN-based descriptors, DeSCA requires no training data for learning convolutional kernels, since the convolutions are defined as the local self-similarity between pairs of image patches, which yields its robustness to cross-modal imaging. Fig. 1 illustrates the robustness of DeSCA for image pairs across non-rigid deformations and illumination changes. In the experimental results, we show that DeSCA outperforms existing area-based and feature-based descriptors on various benchmarks.

2 Related Work

Feature Descriptors Conventional gradient-based descriptors, such as SIFT [14] and DAISY [23], as well as intensity comparison-based binary descriptors, such as BRIEF [24], have shown limited performance in dense correspondence estimation between cross-modal image pairs. Besides these handcrafted features, several attempts have been made using machine learning algorithms to derive features from large-scale datasets [15,25]. A few of these methods use deep convolutional neural networks (CNNs) [26], which have revolutionized image-level classification, to learn discriminative descriptors for local patches. For designing explicit feature descriptors based on a CNN architecture, immediate activations are extracted as the descriptor [15,16,17,18,19], and have been shown to be effective for this patch-level task. However, even though CNN-based descriptors encode a discriminative structure with a deep architecture, they have inherent limitations in cross-modal image correspondence because they are derived from convolutional layers using shared patches or volumes [19,20]. Furthermore, they cannot in practice provide dense descriptors in the image domain due to their prohibitively high computational complexity.

To estimate cross-modal correspondences, variants of the SIFT descriptor have been developed [27], but these gradient-based descriptors maintain an inherent limitation similar to SIFT in dealing with image gradients that vary differently between modalities. For illumination invariant correspondences, Wang

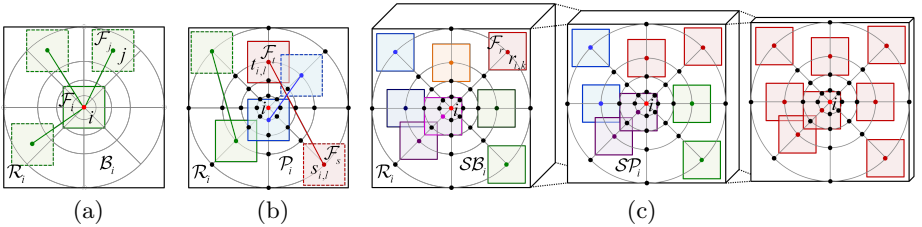


Fig. 2. Illustration of (a) LSS [22] using center-biased dense max pooling, (b) DASC [10] using patch-wise receptive field pooling, and (c) our DeSCA. Boxes, formed by solid and dotted lines, depict source and target patches. DeSCA incorporates a circular spatial pyramid pooling on hierarchical self-convolutional activations.

et al. proposed the local intensity order pattern (LIOP) descriptor [28], but severe radiometric variations may often alter the relative order of pixel intensities. Simo-Serra *et al.* proposed the deformation and light invariant (DaLI) descriptor [29] to provide high resilience to non-rigid image transformations and illumination changes, but it cannot provide dense descriptors in the image domain due to its high computational time.

Schechtman and Irani introduced the LSS descriptor [22] for the purpose of template matching, and achieved impressive results in object detection and retrieval. By employing LSS, many approaches have tried to solve for cross-modal correspondences [30,31,32]. However, none of these approaches scale well to dense matching in cross-modal images due to low discriminative power and high complexity. Inspired by LSS, Kim *et al.* recently proposed the DASC descriptor to estimate cross-modal dense correspondences [10]. Though it can provide satisfactory performance, it is not able to handle non-rigid deformations and has limited discriminative power due to its fixed patch pooling scheme.

Area-Based Similarity Measures A popular measure for registration of cross-modal medical images is mutual information (MI) [33], based on the entropy of the joint probability distribution function, but it provides reliable performance only for variations undergoing a global transformation [34]. Although cross-correlation based methods such as adaptive normalized cross-correlation (ANCC) [35] produce satisfactory results for locally linear variations, they are less effective against more substantial modality variations. Robust selective normalized cross-correlation (RSNCC) [9] was proposed for dense alignment between cross-modal images, but as an intensity based measure it can still be sensitive to cross-modal variations. Recently, DeepMatching [36] was proposed to compute dense correspondences by employing a hierarchical pooling scheme like CNN, but it is not designed to handle cross-modal matching.

3 Background

Let us define an image as $f_i : \mathcal{I} \rightarrow \mathbb{R}$ for pixel i , where $\mathcal{I} \subset \mathbb{N}^2$ is a discrete image domain. Given the image f_i , a dense descriptor $\mathcal{D}_i : \mathcal{I} \rightarrow \mathbb{R}^L$ with a feature dimension of L is defined on a local support window \mathcal{R}_i of size $M_{\mathcal{R}}$.

Unlike conventional descriptors, relying on common visual properties across images such as color and gradient, LSS-based descriptors provide robustness to different imaging modalities since internal self-similarities are preserved across cross-modal image pairs [22,10]. As shown in Fig. 2(a), the LSS discretizes the correlation surface on a log-polar grid, generates a set of bins, and then stores the maximum correlation value of each bin. Formally, it generates an $L^{\text{LSS}} \times 1$ feature vector $\mathcal{D}_i^{\text{LSS}} = \bigcup_l d_i^{\text{LSS}}(l)$ for $l \in \{1, \dots, L^{\text{LSS}}\}$, with $d_i^{\text{LSS}}(l)$ computed as

$$d_i^{\text{LSS}}(l) = \max_{j \in \mathcal{B}_i(l)} \{\exp(-\mathcal{S}(\mathcal{F}_i, \mathcal{F}_j)/\sigma_c)\}, \quad (1)$$

where log-polar bins are defined as $\mathcal{B}_i = \{j | j \in \mathcal{R}_i, \rho_{r-1} < |i - j| \leq \rho_r, \theta_{a-1} < \angle(i - j) \leq \theta_a\}$ with a log radius ρ_r for $r \in \{1, \dots, N_\rho\}$ and a quantized angle θ_a for $a \in \{1, \dots, N_\theta\}$ with $\rho_0 = 0$ and $\theta_0 = 0$. $\mathcal{S}(\mathcal{F}_i, \mathcal{F}_j)$ is a correlation surface between a patch \mathcal{F}_i and \mathcal{F}_j of size $M_{\mathcal{F}}$, computed using sum of square differences. Each pair of r and a is associated with a unique index l . Though LSS provides robustness to modality variations, its significant computation does not scale well for estimating dense correspondences in cross-modal images.

Inspired by the LSS [22], the DASC [10] encodes the similarity between patch-wise receptive fields sampled from a log-polar circular point set \mathcal{P}_i as shown in Fig. 2(b). It is defined such that $\mathcal{P}_i = \{j | j \in \mathcal{R}_i, |i - j| = \rho_r, \angle(i - j) = \theta_a\}$, which has a higher density of points near a center pixel, similar to DAISY [23]. The DASC is encoded with a set of similarities between patch pairs of sampling patterns selected from \mathcal{P}_i such that $\mathcal{D}_i^{\text{DASC}} = \bigcup_l d_i^{\text{DASC}}(l)$ for $l \in \{1, \dots, L^{\text{DASC}}\}$:

$$d_i^{\text{DASC}}(l) = \exp(-(1 - |\mathcal{C}(\mathcal{F}_{s_{i,l}}, \mathcal{F}_{t_{i,l}})|)/\sigma_c), \quad (2)$$

where $s_{i,l}$ and $t_{i,l}$ are the l^{th} selected sampling pattern from \mathcal{P}_i at pixel i . The patch-wise similarity is computed with an exponential function with a bandwidth of σ_c , which has been widely used for robust estimation [37]. $\mathcal{C}(\mathcal{F}_{s_{i,l}}, \mathcal{F}_{t_{i,l}})$ is computed using an adaptive self-correlation measure. While the DASC descriptor has shown satisfactory results for cross-modal dense correspondence [10], its randomized receptive field pooling has limited descriptive power and does not accommodate non-rigid deformations.

4 The DeSCA Descriptor

4.1 Motivation and Overview

Inspired by DASC [10], our DeSCA descriptor also measures an adaptive self-correlation between two patches. We, however, adopt a different strategy for selecting patch pairs, and build self-convolutional activations that more comprehensively encode self-similar structure to improve the discriminative power and the robustness to non-rigid image deformation (Sec. 4.2). Motivated by the deep architecture of CNN-based descriptors [19], we further build hierarchical self-convolution activations to enhance the robustness of the DeSCA descriptor (Sec. 4.4). Densely sampled descriptors are efficiently computed over an entire image using a method based on fast edge-aware filtering (Sec. 4.3). Fig. 2(c) illustrates the DeSCA descriptor, which incorporates a circular spatial pyramid pooling on hierarchical self-convolutional activations.

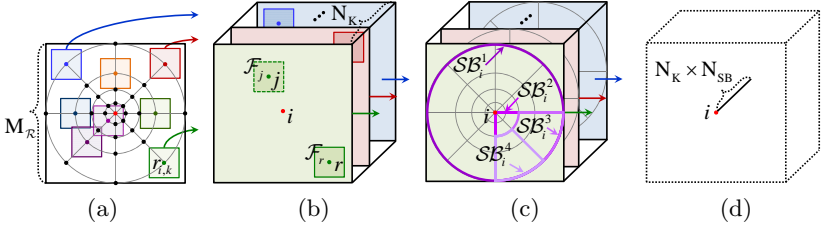


Fig. 3. Computation of single self-convolutional activation (SiSCA). (a) A local support window \mathcal{R}_i of size $M_{\mathcal{R}}^2$ with N_K random samples. (b) For each random patch, a self-convolutional surface is computed using an adaptive self-correlation measure. (c) A self-convolutional activation is then obtained through circular spatial pyramid pooling (C-SPP). (d) The activation from C-SPP is concatenated as 1-D feature vector.

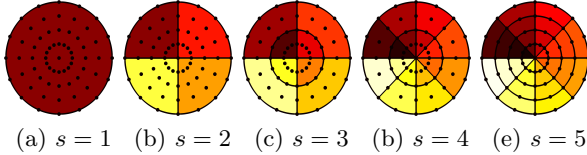


Fig. 4. Examples of the circular spatial pyramidal bins \mathcal{SB}_i . The total number of bins is $N_{\mathcal{SB}} = \sum_{s=2}^{N_S} 2^s + 1$, where N_S represents the pyramid level.

4.2 SiSCA: Single Self-Convolutional Activation

To simultaneously leverage the benefits of self-similarity in DASC [10] and the deep convolutional architecture of CNNs while overcoming the limitations of each method, our approach builds self-convolutional activations. Unlike DASC [10], the feature response is obtained through circular spatial pyramid pooling. We start by describing a single-layer version of DeSCA, which we denote as SiSCA.

Self-Convolutions To build a self-convolutional activation, we randomly select N_K points from a log-polar circular point set \mathcal{P}_i defined within a local support window \mathcal{R}_i . We convolve a patch $\mathcal{F}_{r_{i,k}}$ centered at the k -th point $r_{i,k}$ with all patches \mathcal{F}_j , which is defined for $j \in \mathcal{R}_i$ and $k \in \{1, \dots, N_K\}$ as Fig. 3(b). Similar to DASC [10], the similarity $\mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j)$ between patch pairs is measured using an adaptive self-correlation, which is known to be effective in addressing cross-modality. With (i, k) omitted for simplicity, $\mathcal{C}(\mathcal{F}_r, \mathcal{F}_j)$ is computed as follows:

$$\mathcal{C}(\mathcal{F}_r, \mathcal{F}_j) = \frac{\sum_{r', j'} \omega_{r, r'} (f_{r'} - \mathcal{G}_{r, r}) (f_{j'} - \mathcal{G}_{r, j})}{\sqrt{\sum_{r'} \omega_{r, r'} (f_{r'} - \mathcal{G}_{r, r})} \sqrt{\sum_{j'} \omega_{r, r'} (f_{j'} - \mathcal{G}_{r, j})}}, \quad (3)$$

for $r' \in \mathcal{F}_r$ and $j' \in \mathcal{F}_j$. $\mathcal{G}_{r, r} = \sum_{r'} \omega_{r, r'} f_{r'}$ and $\mathcal{G}_{r, j} = \sum_{j'} \omega_{r, r'} f_{j'}$ represent weighted averages of $f_{r'} \in \mathcal{F}_r$ and $f_{j'} \in \mathcal{F}_j$. Similar to DASC [10], the weight $\omega_{r, r'}$ represents how similar two pixels r and r' are, and is normalized, *i.e.*, $\sum_{r'} \omega_{r, r'} = 1$. It may be defined using any form of edge-aware weighting [38, 39].

Circular Spatial Pyramid Pooling To encode the feature responses on the self-convolutional surface, we propose a circular spatial pyramid pooling (C-SPP)

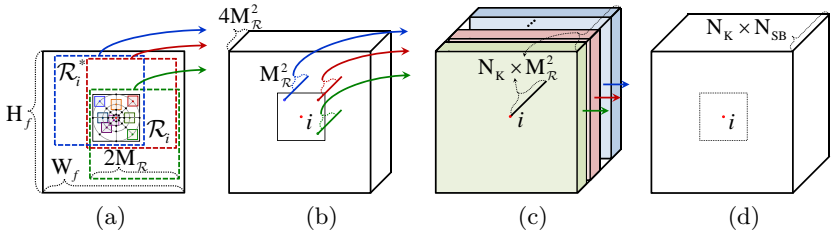


Fig. 5. Efficient computation of self-convolutional activations on the image. (a) An image f_i with a doubled support window R_i^* and random samples. (b) 1-D vectorial self-convolutional surface. (c) Self-convolutional activations. (d) Activations after C-SPP. With an efficient edge-aware filtering and activation reformulation, self-convolutional activations are computed efficiently in a dense manner.

scheme, which pools the responses within each hierarchical spatial bin, similar to a spatial pyramid pooling (SPP) [20,40,41] but in a circular configuration. Note that many existing descriptors also adopt a circular pooling scheme thanks to its robustness based on a higher pixel density near a central pixel [22,23,24]. We further encodes more structure information with a C-SPP.

The circular pyramidal bins $\mathcal{SB}_i(u)$ are defined from log-polar circular bins \mathcal{B}_i , where u indexes all pyramidal level $s \in \{1, \dots, N_S\}$ and all bins in each level s as in Fig. 4. The circular pyramidal bin at the top of pyramid, *i.e.*, $s = 1$, first encompasses all of bins \mathcal{B}_i . At the second level, *i.e.*, $s = 2$, it is defined by dividing \mathcal{B}_i into quadrants. For further lower pyramid levels, *i.e.*, $s > 2$, the circular pyramidal bins are defined differently according to whether s is odd or even. For an odd s , the bins are defined by dividing bins in upper level into two parts along the radius. For an even s , they are defined by dividing bins in upper level into two parts with respect to the angle. The set of all circular pyramidal bins \mathcal{SB}_i is denoted such that $\mathcal{SB}_i = \bigcup_u \mathcal{SB}_i(u)$ for $u \in \{1, \dots, N_{SB}\}$, where the number of circular spatial pyramid bins is defined as $N_{SB} = \sum_{s=2}^{N_S} 2^s + 1$.

As illustrated in Fig. 3(c), the feature responses are finally max-pooled on the circular pyramidal bins $\mathcal{SB}_i(u)$ of each self-convolutional surface $\mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j)$, yielding a feature response

$$h_i(k, u) = \max_{j \in \mathcal{SB}_i(u)} \{\mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j)\}, \quad u \in \{1, \dots, N_{SB}\}. \quad (4)$$

This pooling is repeated for all $k \in \{1, \dots, N_K\}$, yielding accumulated activations $\hat{h}_i(l) = \bigcup_{\{k,u\}} h_i(k, u)$ where l indexes for all k and u .

Interestingly, LSS [22] also uses the max pooling strategy to mitigate the effects of non-rigid image deformation. However, max pooling in the 2-D self-correlation surface of LSS [22] loses fine-scale matching details as reported in [10]. By contrast, DeSCA employs circular spatial pyramid pooling in the 3-D self-correlation surface that provides a more discriminative representation of self-similarities, thus maintaining fine-scale matching details as well as providing robustness to non-rigid image deformations.

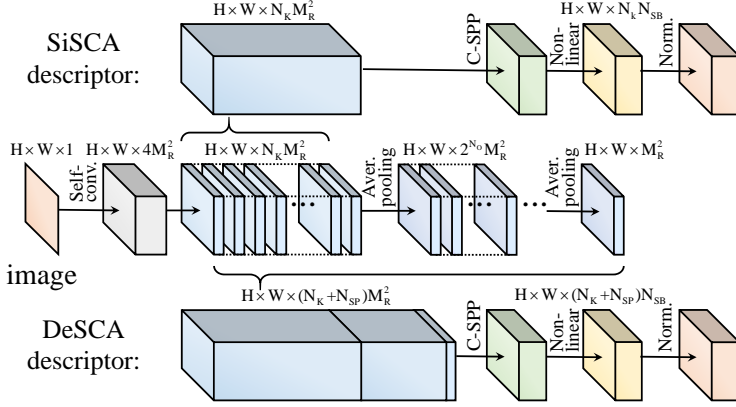


Fig. 6. Visualization of SiSCA and DeSCA descriptor. Our architecture consists of a hierarchical self-convolutional layer, circular spatial pyramid pooling layer, non-linear gating layer, and normalization layer.

Non-linear Gating and Nomalization The final feature responses are passed through a non-linear and normalization layer to mitigate the effects of outliers. With accumulated activations \hat{h}_i , the single self-convolution activation (SiSCA) descriptor $\mathcal{D}_i^{\text{SiSCA}} = \bigcup_l d_i^{\text{SiSCA}}(l)$ is computed for $l \in \{1, \dots, L^{\text{SiSCA}}\}$ through a non-linear gating layer:

$$d_i^{\text{SiSCA}}(l) = \exp(-(1 - |\hat{h}_i(l)|)/\sigma_c), \quad (5)$$

where σ_c is a Gaussian kernel bandwidth. The size of features obtained from the SiSCA becomes $L^{\text{SiSCA}} = N_K N_{SB}$. Finally, $d_i^{\text{SiSCA}}(l)$ for each pixel i is normalized with an L-2 norm for all l .

4.3 Efficient Computation for Dense Description

The most time-consuming part of DeSCA is in constructing self-convolutional surfaces $\mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j)$ for k and j , where $N_K M_R^2$ computations of (3) are needed for each pixel i . Straightforward computation of a weighted summation using ω in (3) would require considerable processing with a computational complexity of $O(IM_{\mathcal{F}} N_K M_R^2)$, where $I = H_f W_f$ represents the image size (height H_f and width W_f). To expedite processing, we utilize fast edge-aware filtering [38,39] and propose a pre-computation scheme for convolutional surfaces.

Similar to DASC [10], we compute $\mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j)$ efficiently by first rearranging the sampling patterns $(r_{i,k}, j)$ into reference-biased pairs $(i, j_r) = (i, i + r_{i,k} - j)$. $\mathcal{C}(\mathcal{F}_i, \mathcal{F}_{j_r})$ can then be expressed as

$$\mathcal{C}(\mathcal{F}_i, \mathcal{F}_{j_r}) = \frac{\mathcal{G}_{i,j_r} - \mathcal{G}_{i,i} \cdot \mathcal{G}_{i,j_r}}{\sqrt{\mathcal{G}_{i,i^2} - (\mathcal{G}_{i,i})^2} \cdot \sqrt{\mathcal{G}_{i,j_r^2} - (\mathcal{G}_{i,j_r})^2}}, \quad (6)$$

where $\mathcal{G}_{i,j_r} = \sum_{i',j'_r} \omega_{i,i'} f_{i'} f_{j'_r}$, $\mathcal{G}_{i,j_r^2} = \sum_{i',j'_r} \omega_{i,i'} f_{j'_r}^2$, and $\mathcal{G}_{i,i^2} = \sum_{i'} \omega_{i,i'} f_{i'}^2$. $\mathcal{C}(\mathcal{F}_i, \mathcal{F}_{j_r})$ can be efficiently computed using any form of fast edge-aware filter

Algorithm 1: Deep Self-Convolutional Activations (DeSCA) Descriptor**Input :** image f_i , random samples $r_{i,k}$.**Output :** DeSCA descriptor $\mathcal{D}_i^{\text{DeSCA}}$.

-
- 1 : Compute $\mathcal{C}(\mathcal{F}_i, \mathcal{F}_j)$ for a doubled support window \mathcal{R}_i^* by using (6).
 - 2 : Estimate $\mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j)$ from $\mathcal{C}(\mathcal{F}_i, \mathcal{F}_j)$ according to the index mapping process.
 for $v = 1 : N_{\mathcal{SP}}$ **do** /* **hierarchical aggregation using average pooling** */
 - 3 : Determine a circular pyramidal point $\mathcal{SP}_i(v)$.
 - 4 : Compute $\mathcal{C}(\mathcal{F}_v, \mathcal{F}_j)$ by using an average pooling for $\mathcal{SP}_i(v)$ on $\mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j)$.
 end for
 - for** $u = 1 : N_{\mathcal{SB}}$ **do** /* **hierarchical spatial aggregation using C-SPP** */
 - 6 : Determine a circular pyramidal bin $\mathcal{SB}_i(u)$.
 - 7 : Compute $h_i(k, u)$ and $h_i(v, u)$ by using C-SPP on each $\mathcal{SB}_i(u)$
 from $\mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j)$ and $\mathcal{C}(\mathcal{F}_v, \mathcal{F}_j)$, respectively.
 end for
 - 8 : Build hierarchical self-convolutional activations $\hat{h}_i(l)$ from $h_i(k, u)$ and $h_i(v, u)$.
 - 8 : Compute a nonlinear response (5), followed by L-2 normalization.
 - 9 : Build a DeSCA descriptor $\mathcal{D}_i^{\text{DeSCA}} = \bigcup_l d_i^{\text{DeSCA}}(l)$.
-

[38,39] with the complexity of $O(IN_K M_{\mathcal{R}}^2)$. $\mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j)$ is then simply obtained from $\mathcal{C}(\mathcal{F}_i, \mathcal{F}_{j_r})$ by re-indexing sampling patterns.

Though we remove the computational dependency on patch size $M_{\mathcal{F}}$, $N_K M_{\mathcal{R}}^2$ computations of (6) are still needed to obtain the self-convolutional activations, where many sampling pairs are repeated. To avoid such redundancy, we first compute self-convolutional activation $\mathcal{C}(\mathcal{F}_i, \mathcal{F}_j)$ for $j \in \mathcal{R}_i^*$ with a doubled local support window \mathcal{R}_i^* of size $2M_{\mathcal{R}} \times 2M_{\mathcal{R}} (= 4M_{\mathcal{R}}^2)$. A doubled local support window is used because (6) is computed with patch \mathcal{F}_{j_r} and the minimum support window size for \mathcal{R}_i^* to cover all samples within \mathcal{R}_i is $2M_{\mathcal{R}}$ as shown in Fig. 5(b). After the self-convolutional activation for \mathcal{R}_i^* is computed once over the image domain, $\mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j)$ can be extracted through an index mapping process, where the indexes for $\mathcal{R}_{i-r_{i,k}}$ are estimated from \mathcal{R}_i^* .

4.4 DeSCA: Deep Self-Convolutional Activations

So far, we have discussed how to build the self-convolutional activation on a single level. In this section, we extend this idea by encoding self-similar structures at multiple levels in a manner similar to a deep architecture widely adopted in the CNNs [26]. DeSCA is defined similarly to SiSCA, except that an average pooling is executed before C-SPP (see Fig. 6). With self-convolutional activations, we perform the average pooling on circular pyramidal point sets.

In comparison to the self-convolutions just from a single patch, the spatial aggregation of self-convolutional responses is clearly more robust, and it requires only marginal computational overhead over SiSCA. The strength of such a hierarchical aggregation has also been shown in [36]. Compared to using only last CNN layer activations, we use all intermediate activations from hierarchical average pooling, which yields better cross-modal matching quality.

To build the hierarchical self-convolutional volume using an average pooling, we first define the circular pyramidal point sets $\mathcal{SP}_i(v)$ from log-polar circular

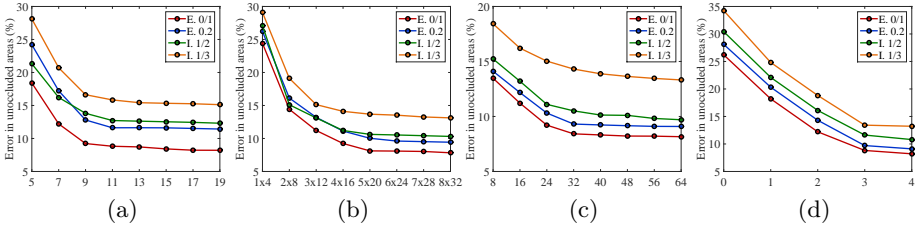


Fig. 7. Component analysis of DeSCA on the Middlebury benchmark [42] for varying parameter values, such as (a) support window size $M_{\mathcal{R}}$, (b) number of log-polar circular point $N_{\rho} \times N_{\theta}$, (c) number of random samples N_K , and (d) level of circular spatial pyramid N_S . In each experiment, all other parameters are fixed to the initial values.

point sets \mathcal{P}_i , where v associates all pyramidal level $o \in \{1, \dots, N_O\}$ and all points in each level o . In the average pooling, the circular pyramidal bins $\mathcal{SB}_i(u)$ used in C-SPP is re-used such that $\mathcal{SP}_i(v) = \{j | j \in \mathcal{P}_i, j \in \mathcal{SB}_i(u)\}$, thus $N_S = N_O$. Deep self-convolutional activations are defined by aggregating $\mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j)$ for all $r_{i,k}$ patches determined on each $\mathcal{SP}_i(v)$ such that

$$\mathcal{C}(\mathcal{F}_v, \mathcal{F}_j) = \sum_{r_{i,k} \in \mathcal{SP}_i(v)} \mathcal{C}(\mathcal{F}_{r_{i,k}}, \mathcal{F}_j) / N_v, \quad (7)$$

which is defined for all v , and N_v is the number of $r_{i,k}$ patches within $\mathcal{SP}_i(v)$. The hierarchical activations are sequentially aggregated using average pooling from bottom to top of circular pyramidal point set $\mathcal{SP}_i(v)$. After computing hierarchical self-convolutional aggregations, similar to SiSCA, the DeSCA employs C-SPP, non-linear, and normalization layer presented in Sec. 4.2. Hierarchical self-convolutional activation $h_i(v, u)$ is computed using the C-SPP such that

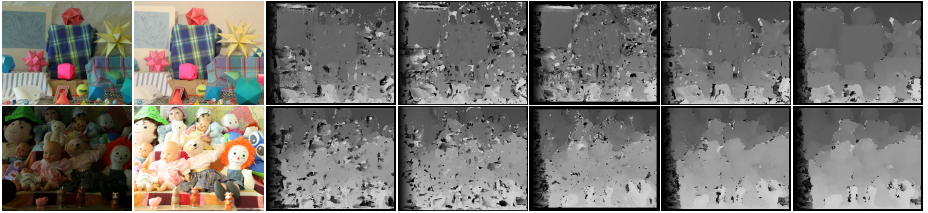
$$h_i(v, u) = \max_{j \in \mathcal{SB}_i(u)} \{\mathcal{C}(\mathcal{F}_v, \mathcal{F}_j)\}. \quad (8)$$

Accumulated self-convolutional activations are built from $h_i(k, u)$ in (4) and $h_i(v, u)$ in (8) such that $\hat{h}_i(l) = \bigcup_{\{k,v,u\}} \{h_i(k, u), h_i(v, u)\}$. Our DeSCA descriptor $d_i^{\text{DeSCA}}(l)$ is then passed through a non-linear layer. $\mathcal{D}_i^{\text{DeSCA}} = \bigcup_l d_i^{\text{DeSCA}}(l)$ is built for $l \in \{1, \dots, L^{\text{DeSCA}}\}$ with $L^{\text{DeSCA}} = (N_K + N_{\mathcal{SP}})N_{\mathcal{SB}}$. Finally, $d_i^{\text{DeSCA}}(l)$ for each pixel i is normalized with an L-2 norm for all l .

5 Experimental Results and Discussion

5.1 Experimental Settings

In our experiments, the DeSCA descriptor was implemented with the following fixed parameter settings for all datasets: $\{\sigma_c, M_{\mathcal{F}}, M_{\mathcal{R}}, N_K, N_S\} = \{0.5, 5, 9, 32, 3\}$, and $\{N_{\rho}, N_{\theta}\} = \{4, 16\}$. We chose the guided filter (GF) for edge-aware filtering in (6), with a smoothness parameter of $\epsilon = 0.03^2$. We implemented the DeSCA descriptor in C++ on an Intel Core i7-3770 CPU at 3.40 GHz. We will make our code publicly available. The DeSCA descriptor was compared to other state-of-the-art descriptors (SIFT [14], DAISY [23], BRIEF [24], LIOP [28], DaLI [29],



(a) image 1 (b) image 2 (c) ANCC (d) SIFT (e) LSS (f) DASC (g) DeSCA

Fig. 8. Comparison of disparity estimations for *Moebius* and *Dolls* image pairs across illumination combination ‘1/3’ and exposure combination ‘0/2’, respectively. Compared to other methods, DeSCA estimates more accurate and edge-preserved disparity maps.

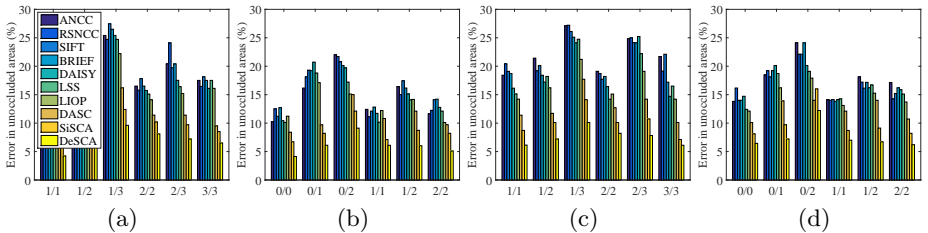


Fig. 9. Average bad-pixel error rate on the Middlebury benchmark [42] with illumination and exposure variations. Optimization was done by GC in (a), (b), and by WTA in (c), (d). DeSCA descriptor shows the best performance with the lowest error rate.

LSS [22], and DASC [10]), as well as area-based approaches (ANCC [35] and RSNC [9]). Furthermore, to evaluate the performance gain with a deep architecture, we compared SiSCA and DeSCA.

5.2 Parameter Evaluation

The matching performance of DeSCA is exhibited in Fig. 7 for varying parameter values, including support window size $M_{\mathcal{R}}$, number of log-polar circular points $N_{\rho} \times N_{\theta}$, number of random samples N_K , and levels of the circular spatial pyramid N_S . Note that $N_O = N_S$. Especially, Fig. 7(c), (d) prove the effectiveness of self-convolutional activations and deep architectures of DeSCA. For a quantitative analysis, we measured the average bad-pixel error rate on the Middlebury benchmark [42]. With a larger support window $M_{\mathcal{R}}$, the matching quality improves rapidly until about 9×9 . $N_{\rho} \times N_{\theta}$ influences the performance of circular pooling, which is found to plateau at 4×16 . Using a larger number of random samples N_K yields better performance since the descriptor encodes more information. The level of circular spatial pyramid N_S also affects the amount of encoding in DeSCA. Based on these experiments, we set $N_K = 32$ and $N_S = 3$ in consideration of efficiency and robustness.

5.3 Middlebury Stereo Benchmark

We evaluated DeSCA on the Middlebury stereo benchmark [42], which contains illumination and exposure variations. In the experiments, the illumination (exposure) combination ‘1/3’ indicates that two images were captured under the



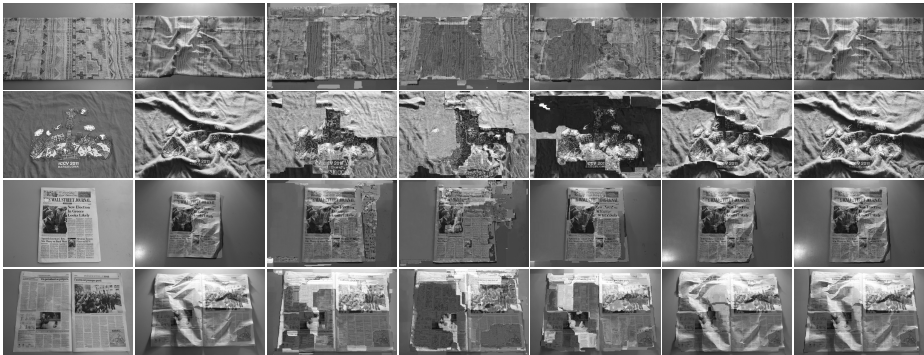
(a) image 1 (b) image 2 (c) BRIEF (d) LSS (e) DASC (f) SiSCA (g) DeSCA
Fig. 10. Dense correspondence evaluations for (from top to bottom) RGB-NIR, flash-noflash, different exposures, and blurred-sharp images. Compared to others, DeSCA estimates more reliable dense correspondences for challenging cross-modal pairs.

Methods	WTA optimization				SF optimization [11]			
	RGB-NIR	flash-noflash	diff. expo.	blur-sharp	RGB-NIR	flash-noflash	diff. expo.	blur-sharp
ANCC [35]	23.21	20.42	25.19	26.14	18.45	14.14	11.96	19.24
RSNCC [9]	27.51	25.12	18.21	27.91	13.41	15.87	9.15	18.21
SIFT [14]	24.11	18.72	19.42	27.18	18.51	11.06	14.87	20.78
DAISY [23]	27.61	26.30	20.72	27.41	20.42	10.84	12.71	22.91
BRIEF [24]	29.14	18.29	17.13	26.43	17.54	9.21	9.54	19.72
LSS [22]	27.82	19.18	18.21	26.14	16.14	11.88	9.11	18.51
LIOP [28]	24.42	16.42	14.22	20.42	15.32	11.42	10.22	17.12
DASC [10]	14.51	13.24	10.32	16.42	13.42	7.11	7.21	11.21
SiSCA	10.12	10.12	8.22	14.22	9.12	6.18	5.22	9.12
DeSCA	8.12	8.22	6.72	13.28	7.62	5.12	4.72	8.01

Table 1. Comparison of quantitative evaluation on cross-modal benchmark.

1^{st} and 3^{rd} illumination (exposure) conditions. For a quantitative evaluation, we measured the bad-pixel error rate in non-occluded areas of disparity maps [42].

Fig. 8 shows the disparity maps estimated under severe illumination and exposure variations with winner-takes-all (WTA) optimization. Fig. 9 displays the average bad-pixel error rates of disparity maps obtained under illumination or exposure variations, with graph-cut (GC) [43] and WTA optimization. Area-based approaches (ANCC [35] and RSNCC [9]) are sensitive to severe radiometric variations, especially when local variations occur frequently. Feature descriptor-based methods (SIFT [14], DAISY [23], BRIEF [24], LSS [22], and DASC [10]) perform better than the area-based approaches, but they also provide limited performance. Our DeSCA descriptor achieves the best results both quantitatively and qualitatively. Compared to SiSCA descriptor, the performance of DeSCA descriptor is highly improved, where the performance benefits of the deep architecture are apparent.



(a) image 1 (b) image 2 (c) DAISY (d) BRIEF (e) LSS (f) DaLI (g) DeSCA

Fig. 11. Dense correspondence comparisons for images with different illumination conditions and non-rigid image deformations [29]. Compared to other approaches, DeSCA provides more accurate dense correspondence estimates with reduced artifacts.

Methods	def.	illum.	def./ illum.	aver.
SIFT [14]	45.15	40.81	47.51	44.49
DAISY [23]	43.98	42.72	43.42	43.37
BRIEF [24]	41.51	37.14	41.35	40
LSS [22]	40.81	39.54	40.11	40.12
LIOP [28]	28.72	31.72	30.21	30.22
DaLI [29]	27.12	27.31	27.99	27.47
DASC [10]	26.21	24.83	27.51	26.18
SiSCA	23.42	22.21	24.17	23.27
DeSCA	20.14	20.72	21.87	20.91

Table 2. Average error rates on the DaLI benchmark.

5.4 Cross-modal and Cross-spectral Benchmark

We evaluated DeSCA on a cross-modal and cross-spectral benchmark [10] containing various kinds of image pairs, namely RGB-NIR, different exposures, flash-noflash, and blurred-sharp. Optimization for all descriptors and similarity measures was done using WTA and SIFT flow (SF) with hierarchical dual-layer belief propagation [11], for which the code is publicly available. Sparse ground truths for those images are used for error measurement as done in [10].

Fig. 10 provides a qualitative comparison of the DeSCA descriptor to other state-of-the-art approaches. As already described in the literature [9], gradient-based approaches such as SIFT [14] and DAISY [23] have shown limited performance for RGB-NIR pairs where gradient reversals and inversions frequently appear. BRIEF [24] cannot deal with noisy regions and modality-based appearance differences since it is formulated on pixel differences only. Unlike these approaches, LSS [22] and DASC [10] consider local self-similarities, but LSS is lacking in discriminative power for dense matching. DASC also exhibits limited performance. Compared to those methods, the DeSCA displays better corre-

image size	SIFT	DAISY	LSS	DaLI	DASC	DeSCA*	DeSCA†
463×370	130.3s	2.5s	31s	352.2s	2.7s	193.2s	9.2s

Table 3. Computation speed of DeSCA and other state-of-the-art local and global descriptors. The brute-force and efficient implementations of DeSCA are denoted by * and †, respectively.

spondence estimation. We also performed a quantitative evaluation with results listed in Table 1, which also clearly demonstrates the effectiveness of DeSCA.

5.5 DaLI Benchmark

We also evaluated DeSCA on a recent, publicly available dataset featuring challenging non-rigid deformations and very severe illumination changes [29]. Fig. 11 presents dense correspondence estimates for this benchmark [29]. A quantitative evaluation is given in Table 2 using ground truth feature points sparsely extracted for each image, although DeSCA is designed to estimate dense correspondences. As expected, conventional gradient-based and intensity comparison-based feature descriptors, including SIFT [14], DAISY [23], and BRIEF [24], do not provide reliable correspondence performance. LSS [22] and DASC [10] exhibit relatively high performance for illumination changes, but are limited on non-rigid deformations. LIOP [28] provides robustness to radiometric variations, but is sensitive to non-rigid deformations. Although DaLI [29] provides robust correspondences, it requires considerable computation for dense matching. DeSCA offers greater discriminative power as well as more robustness to non-rigid deformations in comparison to the state-of-the-art cross-modality descriptors.

5.6 Computational Speed

In Table 3, we compared the computational speed of DeSCA to state-of-the-art local descriptor, namely DaLI [29], and dense descriptors, namely DAISY [23], LSS [22], and DASC [10]. Even though DeSCA needs more computational time compared to some previous dense descriptors, it provides significantly improved matching performance as described previously.

6 Conclusion

The deep self-convolutional activations (DeSCA) descriptor was proposed for establishing dense correspondences between images taken under different imaging modalities. Its high performance in comparison to state-of-the-art cross-modality descriptors can be attributed to its greater robustness to non-rigid deformations because of its effective pooling scheme, and more importantly its heightened discriminative power from a more comprehensive representation of self-similar structure and its formulation in a deep architecture. DeSCA was validated on an extensive set of experiments that cover a broad range of cross-modal differences. In future work, thanks to the robustness to non-rigid deformations and high discriminative power, DeSCA can potentially benefit object detection and semantic segmentation.

References

1. Brown, M., Susstrunk, S.: Multispectral sift for scene category recognition. In: CVPR (2011)
2. Yan, Q., Shen, X., Xu, L., Zhuo, S.: Cross-field joint image restoration via scale map. In: ICCV (2013)
3. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: CVPR (2015)
4. Krishnan, D., Fergus, R.: Dark flash photography. In: SIGGRAPH (2009)
5. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shechtman, E.: Robust patch-based hdr reconstruction of dynamic scenes. In: SIGGRAPH (2012)
6. HaCohen, Y., Shechtman, E., Lishchinski, E.: Deblurring by example using dense correspondence. In: ICCV (2013)
7. Lee, H., Lee, K.: Dense 3d reconstruction from severely blurred images using a single moving camera. In: CVPR (2013)
8. Petschnigg, G., Agrawals, M., Hoppe, H.: Digital photography with flash and no-flash image pairs. In: SIGGRAPH (2004)
9. Shen, X., Xu, L., Zhang, Q., Jia, J.: Multi-modal and multi-spectral registration for natural images. In: ECCV (2014)
10. Kim, S., Min, D., Ham, B., Ryu, S., Do, M.N., Sohn, K.: Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In: CVPR (2015)
11. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. PAMI* **33**(5) (2011) 815–830
12. Kim, J., Liu, C., Sha, F., Grauman, K.: Deformable spatial pyramid matching for fast dense correspondences. In: CVPR (2013)
13. Pinggera, P., Breckon, T., Bischof, H.: On cross-spectral stereo matching using dense gradient features. In: BMVC (2012)
14. Lowe, D.: Distinctive image features from scale-invariant keypoints. *IJCV* **60**(2) (2004) 91–110
15. Simonyan, K., Vedaldi, A., Zisserman, A.: Learning local feature descriptors using convex optimisation. *IEEE Trans. PAMI* **36**(8) (2014) 1573–1585
16. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: ECCV (2014)
17. Fischer, P., Dosovitskiy, A., Brox, T.: Descriptor matching with convolutional neural networks: A comparison to sift. [arXiv:1405.5769](https://arxiv.org/abs/1405.5769) (2014)
18. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: ICML (2014)
19. Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., Moreno-Noguer, F.: Discriminative learning of deep convolutional feature point descriptors. In: ICCV (2015)
20. Dong, J., Soatto, S.: Domain-size pooling in local descriptors: Dsp-sift. In: CVPR (2015)
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
22. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: CVPR (2007)
23. Tola, E., Lepetit, V., Fua, P.: Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Trans. PAMI* **32**(5) (2010) 815–830

24. Calonder, M.: Brief : Computing a local binary descriptor very fast. *IEEE Trans. PAMI* **34**(7) (2011) 1281–1298
25. Trzcinski, T., Christoudias, M., Lepetit, V.: Learning image descriptor with boosting. *IEEE Trans. PAMI* **37**(3) (2015) 597–610
26. Alex, K., Ilya, S., Geoffrey, E.H.: Imagenet classification with deep convolutional neural networks. In: *NIPS* (2012)
27. Saleem, S., Sablatnig, R.: A robust sift descriptor for multispectral images. *IEEE SPL* **21**(4) (2014) 400–403
28. Wang, Z., Fan, B., Wu, F.: Local intensity order pattern for feature description. In: *ICCV* (2011)
29. Simo-Serra, E., Torras, C., Moreno-Noguer, F.: Dali: Deformation and light invariant descriptor. *IJCV* **115**(2) (2015) 136–154
30. Heinrich, P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, V., Brady, S., Schnabel, A.: Mind: Modality indepdent neighbourhood descriptor for multi-modal deformable registration. *MIA* **16**(3) (2012) 1423–1435
31. Torabi, A., Bilodeau, G.: Local self-similarity-based registration of human rois in pairs of stereo thermal-visible videos. *PR* **46**(2) (2013) 578–589
32. Ye, Y., Shan, J.: A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences. *JPRS* **90**(7) (2014) 83–95
33. Pluim, J., Maintz, J., Viergever, M.: Mutual information based registration of medical images: A survey. *IEEE Trans. MI* **22**(8) (2003) 986–1004
34. Heo, Y., Lee, K., Lee, S.: Joint depth map and color consistency estimation for stereo images with different illuminations and cameras. *IEEE Trans. PAMI* **35**(5) (2013) 1094–1106
35. Heo, Y., Lee, K., Lee, S.: Robust stereo matching using adaptive normalized cross-correlation. *IEEE Trans. PAMI* **33**(4) (2011) 807–822
36. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Deepflow: Large displacement optical flow with deep matching. In: *ICCV* (2013)
37. Black, M.J., Sapiro, G., Marimont, D.H., Heeger, D.: Robust anisotropic diffusion. *IEEE Trans. IP* **7**(3) (1998) 421–432
38. Gastal, E., Oliveira, M.: Domain transform for edge-aware image and video processing. In: *SIGGRAPH* (2011)
39. He, K., Sun, J., Tang, X.: Guided image filtering. *IEEE Trans. PAMI* **35**(6) (2013) 1397–1409
40. Seidenari, L., Serra, G., Bagdanov, A.D., Bimbo, A.D.: Local pyramidal descriptors for image recognition. *IEEE Trans. PAMI* **36**(5) (2014) 1033–1040
41. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. PAMI* **37**(9) (2015) 1904–1916
42. online.: <http://vision.middlebury.edu/stereo/>.
43. Boykov, Y., Yeksler, O., Zabih, R.: Fast approximation enermgy minimization via graph cuts. *IEEE Trans. PAMI* **23**(11) (2001) 1222–1239